

UCSF CaBIG: QPACA

Quantitative Pathway Analysis in Cancer



Ajay N. Jain, PhD

Associate Professor, Cancer Research Institute
and Dept. of Laboratory Medicine
University of California, San Francisco

Barbara Novak

BMI PhD student

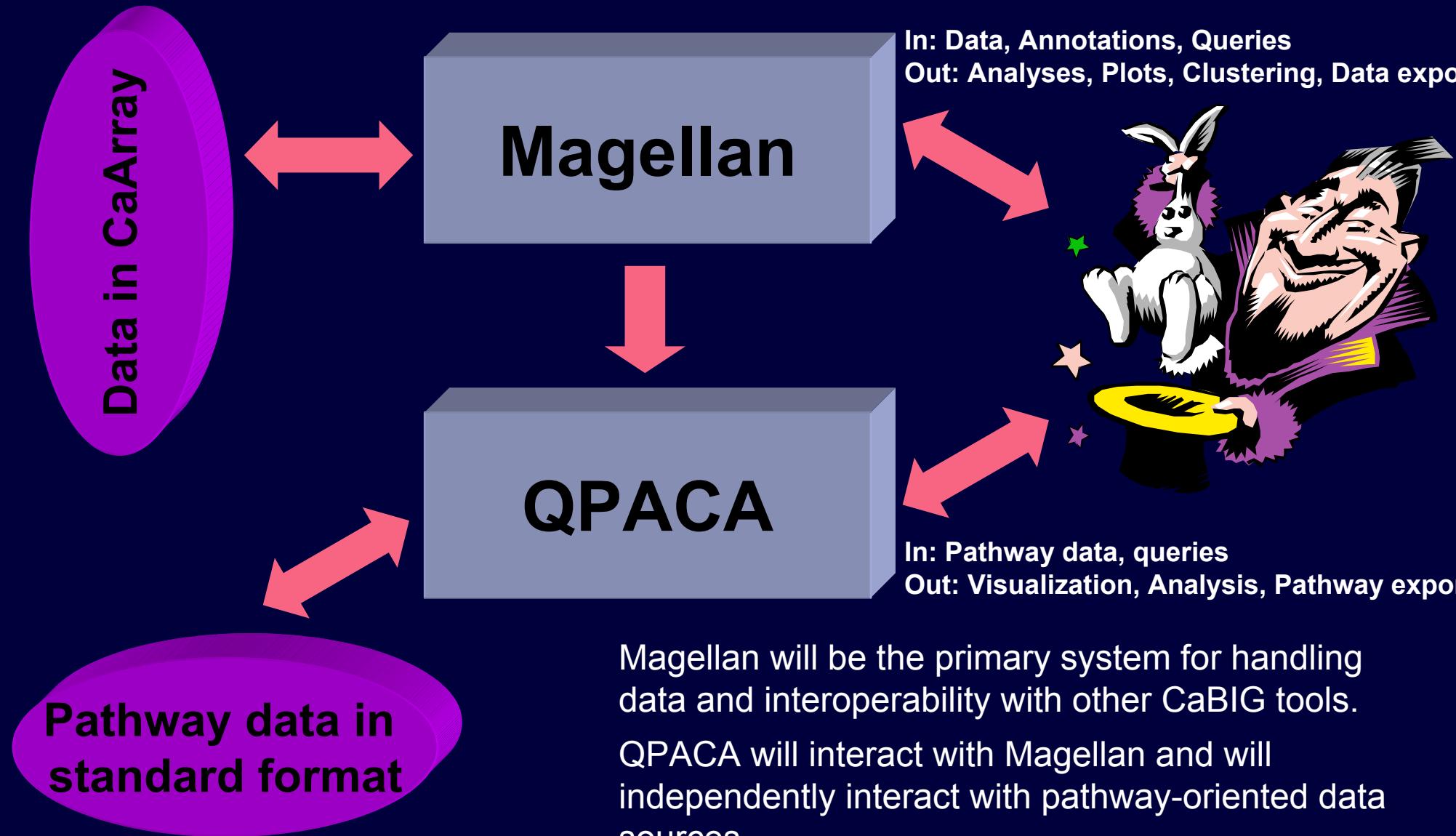
ajain@jainlab.org

<http://www.jainlab.org>

Copyright © 2004, Ajay N. Jain, All Rights Reserved

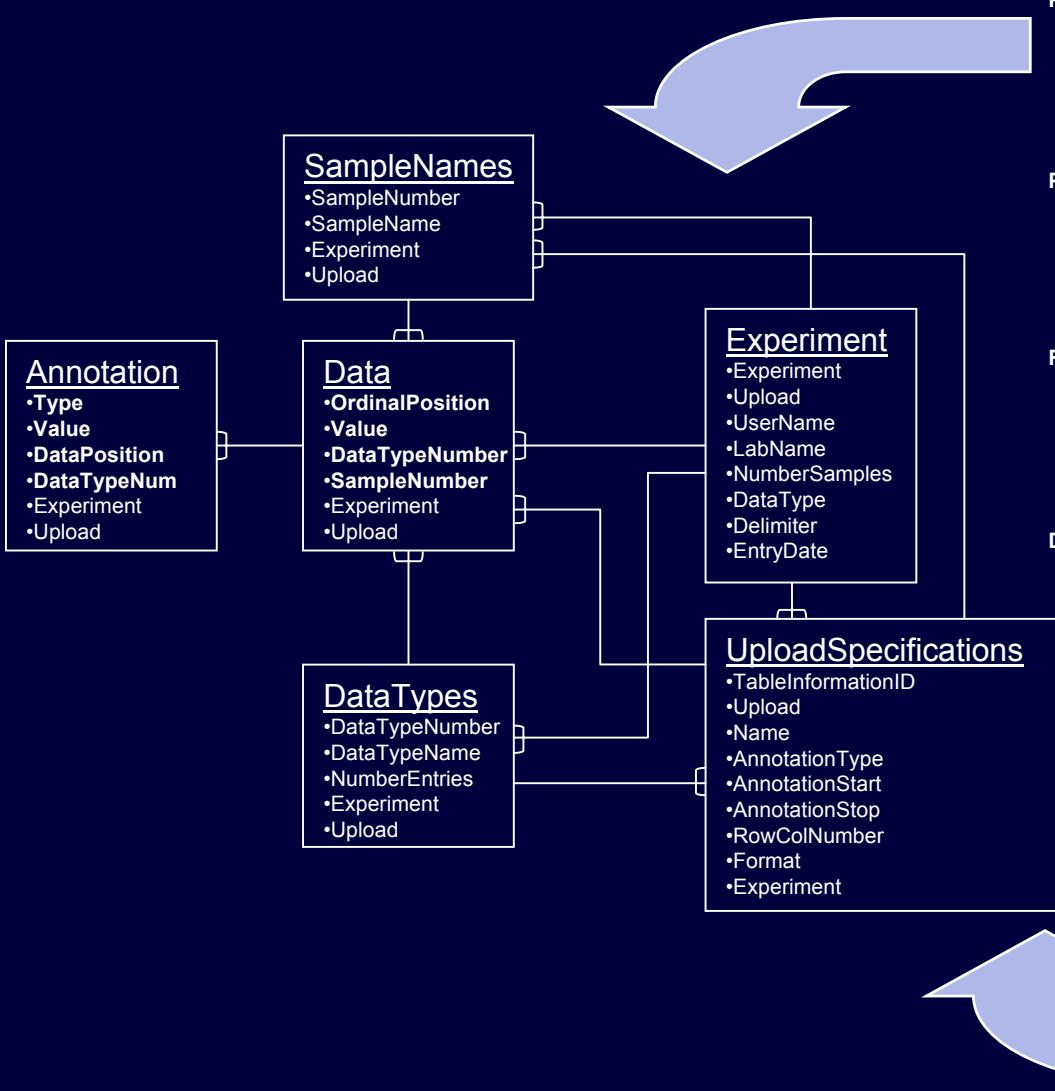


UCSF CaBIG: The BIG picture





Magellan handles experimental + annotation data and provides analytical and visualization methods



Phenotype

Proliferation
Measureable phenotypes.

Apoptosis

Protein

P_1
Protein status for over multiple conditions.

P_n

RNA

G_1
Gene expression levels over multiple conditions.

G_n

DNA

L_1
DNA copy number over the entire genome.

L_n

Cell Lines

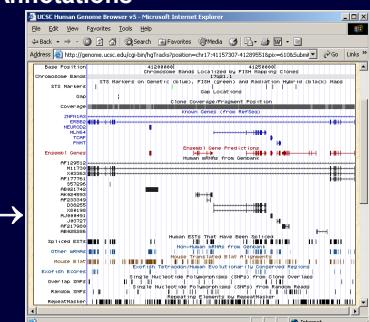
C ₁	□	□	□	□	□	□	...	□
⋮	□	□	□	□	□	□	...	□
C _n	□	□	□	...	□	□	...	□
C ₁	□	□	□	□	□	□	...	□
⋮	□	□	□	□	□	□	...	□
C _n	□	□	□	...	□	□	...	□
C ₁	□	□	□	□	□	□	...	□
⋮	□	□	□	□	□	□	...	□
C _n	□	□	□	...	□	□	...	□
C ₁	□	□	□	□	□	□	...	□
⋮	□	□	□	□	□	□	...	□
C _n	□	□	□	...	□	□	...	□
C ₁	□	□	□	□	□	□	...	□
⋮	□	□	□	□	□	□	...	□
C _n	□	□	□	...	□	□	...	□
C ₁	□	□	□	□	□	□	...	□
⋮	□	□	□	□	□	□	...	□
C _n	□	□	□	...	□	□	...	□
C ₁	□	□	□	□	□	□	...	□
⋮	□	□	□	□	□	□	...	□
C _n	□	□	□	...	□	□	...	□
C ₁	□	□	□	□	□	□	...	□
⋮	□	□	□	□	□	□	...	□
C _n	□	□	□	...	□	□	...	□
C ₁	□	□	□	□	□	□	...	□
⋮	□	□	□	□	□	□	...	□
C _n	□	□	□	...	□	□	...	□

ERBB2:

EC Number: 2.7.1.112

oncogenesis
cell proliferation
Neu/ErbB-2 receptor
protein phosphorylation
protein dephosphorylation
cell growth and maintenance
receptor signaling tyrosine kinase

← Gene Annotations



Genomic Mapping + Context →



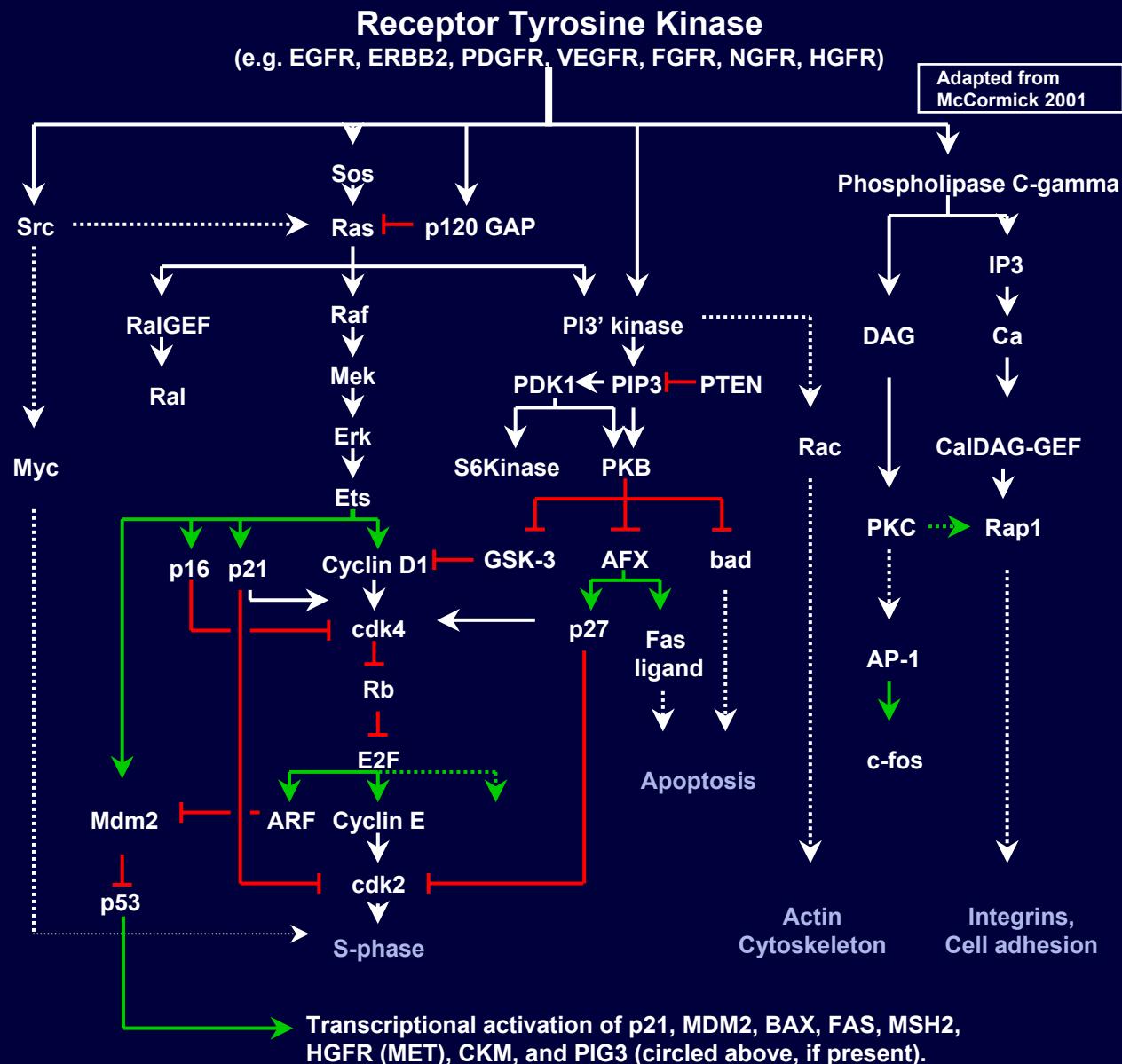
QPACA handles pathway-related questions to support cancer research

Visualization of pathways with data

Projection of statistics onto pathway structures

Inference of gene set relatedness

Inference of biochemical network topology





We represent pathway structure explicitly using a simple language

Informal representations of biochemical pathways must be formalized

Mappings from experimental data space to pathway space are required

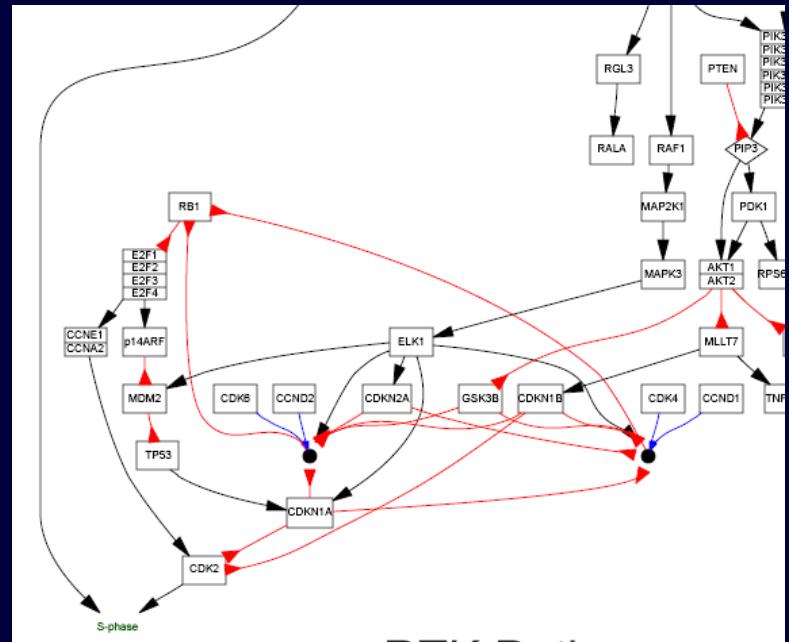
Direct questions involving pathway arms and gene product sets are then easily asked

Pathway Knowledge

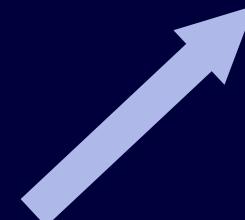


Structured pathway language

```
pathway_name = "RTK Pathway" #element section
pathway_elements {
    cdk4[locusid=1019;type=gene_product]
    cnd1[locusid=595;type=gene_product]
    rb1[locusid=5925;type=gene_product]
    cdk2[locusid=1017;type=gene_product]
    akt1[locusid=208;type=gene_product]
    akt1[locusid=207;type=gene_product]
    ccne1[locusid=898;type=gene_product]
    ccna1[locusid=899;type=gene_product]
    cna2[locusid=890;type=gene_product]
    cdk4_compound[members=cdk4,cnd1,type=compound]
    cdk6_compound[members=cdk6,cnd2,type=compound]
} #pathway section
pathway_elk1 > cdk4_compound elk1 -> cdk2a elk1
pathway_elk1 > cdk4_compound elk1 -> cdk6_compound
elk1 -> cdk6_compound
elk1 -> cdk4_compound [rb1 -> rb1 -> e2f family e2f family -> p14arf e2f family -> cne1 family cne1 family -> cdk2 p14arf -> mdm2 mdm2 -> tp53 tp53 -> cdkn1a cdkn1a -> cdk2 cdkn1a -> cdk6_compound cdkn1a -> cdk4_compound cdkn2a -> cdk6_compound cdkn2a -> cdk4_compound cdkn1b -> cdk2 cdkn1b -> cdk6_compound cdkn1b -> cdk4_compound pip3 -> akt family akt family -> gsk3b gsk3b -> cdk4_compound gsk3b -> cdk6_compound cdk2 -> s-phase }
```



RTK Pathway





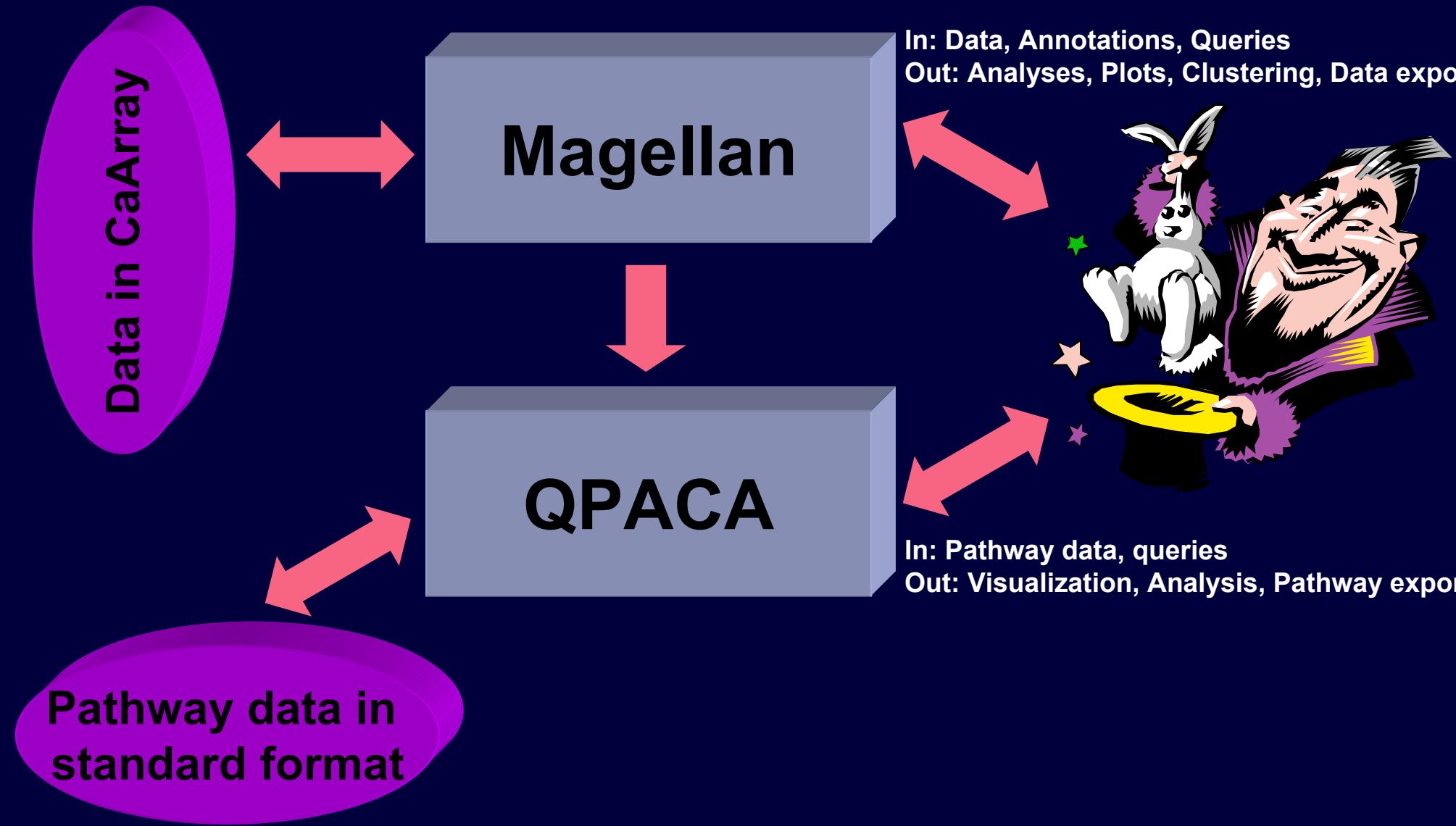
Pathway language is simple, intuitive, and extensible

```
pathway_name = "RTK Pathway"
pathway_elements {
    cdk4[locusid=1019;type=gene_product]
    ccnd1[locusid=595;type=gene_product]
    rb1[locusid=5925;type=gene_product]
    cdk2[locusid=1017;type=gene_product]
    akt2[locusid=208;type=gene_product]
    akt1[locusid=207;type=gene_product]
    ccne1[locusid=898;type=gene_product]
    ccna2[locusid=890;type=gene_product]
    gsk3b[locusid=2932;type=gene_product]
    cdkn1b[locusid=1027;type=gene_product]
    cdk6[locusid=1021;type=gene_product]
    ccnd2[locusid=894;type=gene_product]
    cdkn1a[locusid=1026;type=gene_product]
    s-phase[type=process]
    pip3[type=molecule]
    akt_family[members=akt1,akt2;type=alt]
    cdk4_compound[members=cdk4,ccnd1;type=compound]
    cdk6_compound[members=cdk6,ccnd2;type=compound]
}
```

```
#pathway sections
pathway {
    elk1 -> mdm2
    elk1 -> cdkn2a
    elk1 -> cdkn1a
    elk1 -> cdk4_compound
    elk1 -> cdk6_compound
    cdk6_compound -| rb1
    cdk4_compound -| rb1
    rb1 -| e2f_family
    e2f_family -> p14arf
    e2f_family -> ccne_family
    ccne_family -> cdk2
    p14arf -| mdm2
    mdm2 -| tp53
    tp53 -> cdkn1a
    pip3 -> akt_family
    akt_family -| gsk3b
    gsk3b -| cdk4_compound
    gsk3b -| cdk6_compound
    cdk2 -> s-phase
}
```



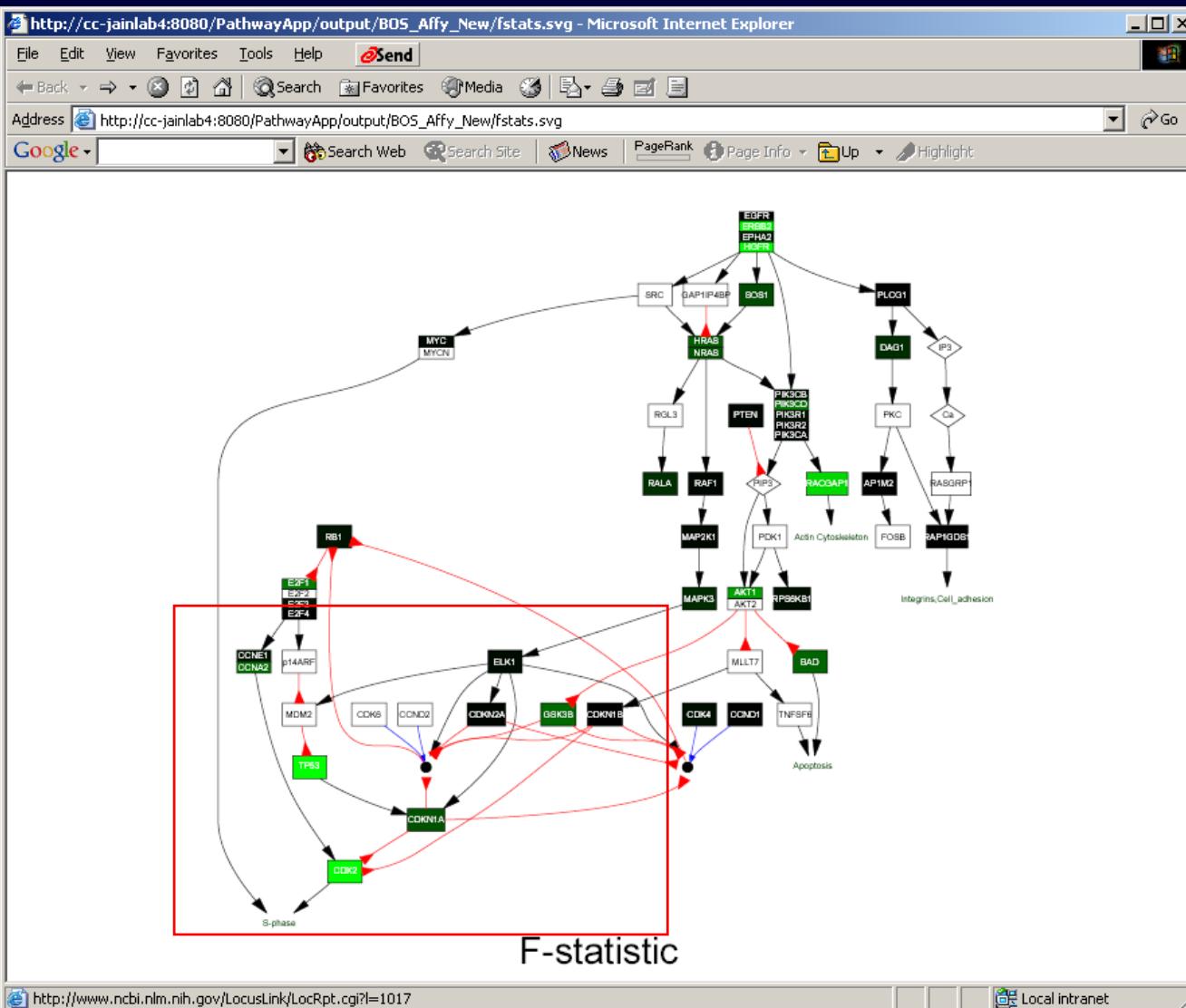
UCSF CaBIG: The BIG picture



CaBIG Goal: Seamless integration with Magellan



We can make use of phenotypes (e.g. ERBB2 status) and compute statistics about expression



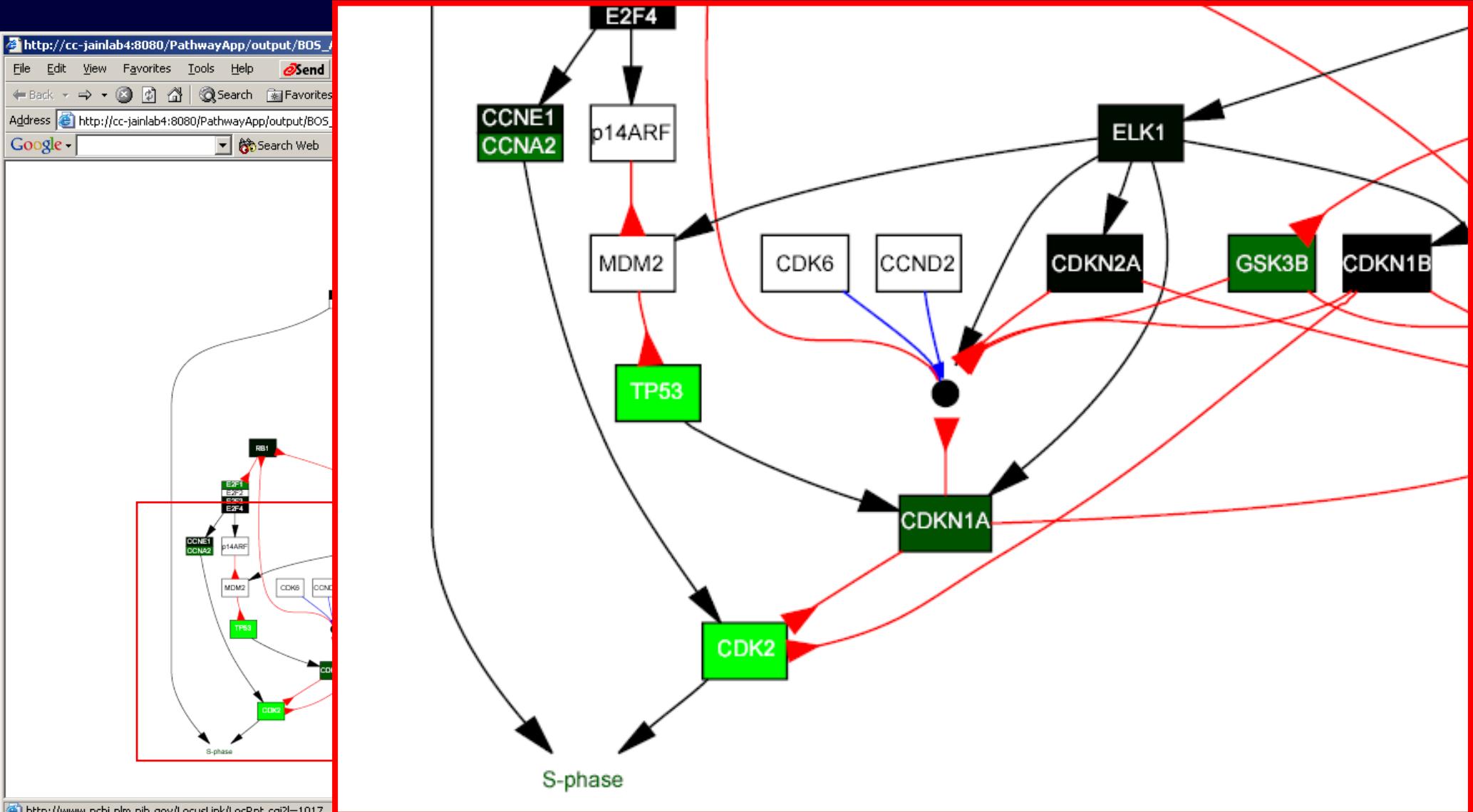
Pathway structure generated from curated representation

Pathway members colored by F-statistic based on class labeling

ERBB2, HGFR, TP53, CDK2, RACGAP1, + others are very different in the two cases (high vs low ERBB2)



The individual nodes can be linked to external annotations





The individual nodes can be linked to external annotations

http://cc-jainlab4:8080/PathwayApp/output/BOS_A

File Edit View Favorites Tools Help Send

Back Search Favorites

Address http://cc-jainlab4:8080/PathwayApp/output/BOS_A

Google Search Web

S-phase

NCBI LocusLink

PubMed Entrez BLAST OMIM Map Viewer Taxonomy Structure

Search LocusLink Display Brief Organism: All

Query: Go Clear

LocusLink Home

CDK2 Index:

- Top of Page
- Nomenclature
- Overview
- Function
- Relationships
- Map
- RefSeq
- Related Seqs
- Links

LocusLink:

- Collaborators
- Download
- FAQ

View Hs CDK2 One of 1 Loci Save All Loci

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Click to Display mRNA-Genomic Alignments (spanning 6009 bps)

Gene	PUB	OMIM	ACEVIEW	UNIGENE	MAP	VAR	HOMOL
GDB	e!	UCSC					

Homo sapiens Official Gene Symbol and Name ([HGNC](#))

CDK2: cyclin-dependent kinase 2

LocusID: 1017

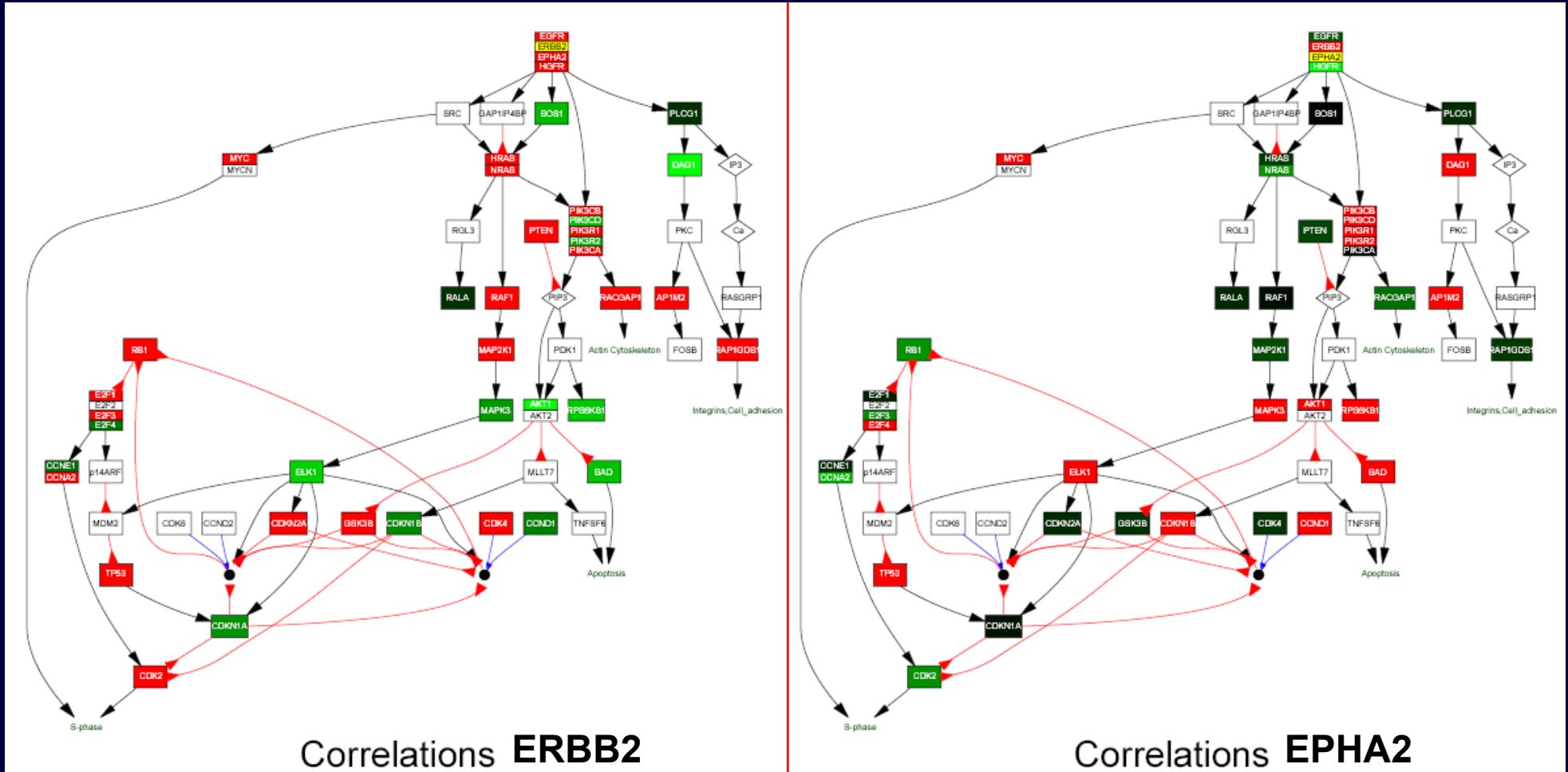
Overview ?

RefSeq Summary: The protein encoded by this gene is a member of the Ser/Thr protein kinase family. This protein kinase is highly similar to the gene products of *S. cerevisiae* cdc28, and *S. pombe* cdc2. It is a catalytic subunit of the cyclin-dependent protein kinase complex, whose activity is

http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=1017



We can compute and visualize gene/gene correlations



The correlations of genes to ERBB2 (left) and EPHA2 (right) are opposite in polarity in many cases. There appears to be a significant switch around EPHA2/ERBB2/ERBB3. Data and biological observations: J. Yeh, L. Timmerman, R. Neve, J. Gray, F. McCormick, J. Gray.



QPACA: Gene set recognition

Input

- ◆ A set of genes hypothesized to be part of a pathway
- ◆ Experimental data (e.g. expression, CGH, proteomic data) with the property that coordination in variation may be related to “pathway-ness”
- ◆ The hypothesis may come from any source

Theoretical issue

- ◆ It is unlikely that all of the experimental samples that are part of the data set are relevant to the study of a *particular* pathway
- ◆ So, we have developed scoring functions and optimization procedures to *select* the samples that are most *relevant*

Output

- ◆ Likelihood that the gene set is part of a pathway or coordinated process

CaBIG Goal: Provide tool to decide “Are these genes in a pathway?”

Organism	Pathway ID	Pathway Name	Number of genes	QPACA p-value	QPACA Score
Human NCI60 Data	hsa04010	MAPK signaling	41	0.01	0.1
	hsa04110	Cell cycle	30	0.01	0.13
	hsa04510	Integrin-mediated cell adhesion	25	0.01	0.16
	hsa00010	Glycolysis / Gluconeogenesis	24	< 0.01	0.25
	hsa04210	Apoptosis	19	0.05	0.18
	hsa00052	Galactose metabolism	13	< 0.01	0.38
	hsa04070	Phosphatidylinositol signaling	13	0.2	0.2
	hsa04620	Toll-like receptor signaling	12	0.05	0.3
	hsa04350	TGF-Beta signaling	11	0.05	0.35
Yeast Hughes Data	sce04110	Cell cycle	100	<< 0.01	0.24
	sce00230	Purine metabolism	95	< 0.01	0.24
	sce04010	MAPK signaling	55	0.01	0.22
	sce00562	Inositol phosphate metabolism	32	0.5	0.1
	sce00052	Galactose metabolism	30	<< 0.01	0.83
	sce04020	Second messenger signaling	19	0.01	0.47
	sce00100	Sterols biosynthesis	15	< 0.01	0.68



QPACA: Edge prediction

Input

- ◆ A set of genes hypothesized to be part of a pathway
- ◆ Experimental data (e.g. expression, CGH, proteomic data) with the property that coordination in variation may be related to “pathway-ness”
- ◆ The hypothesis may come from any source

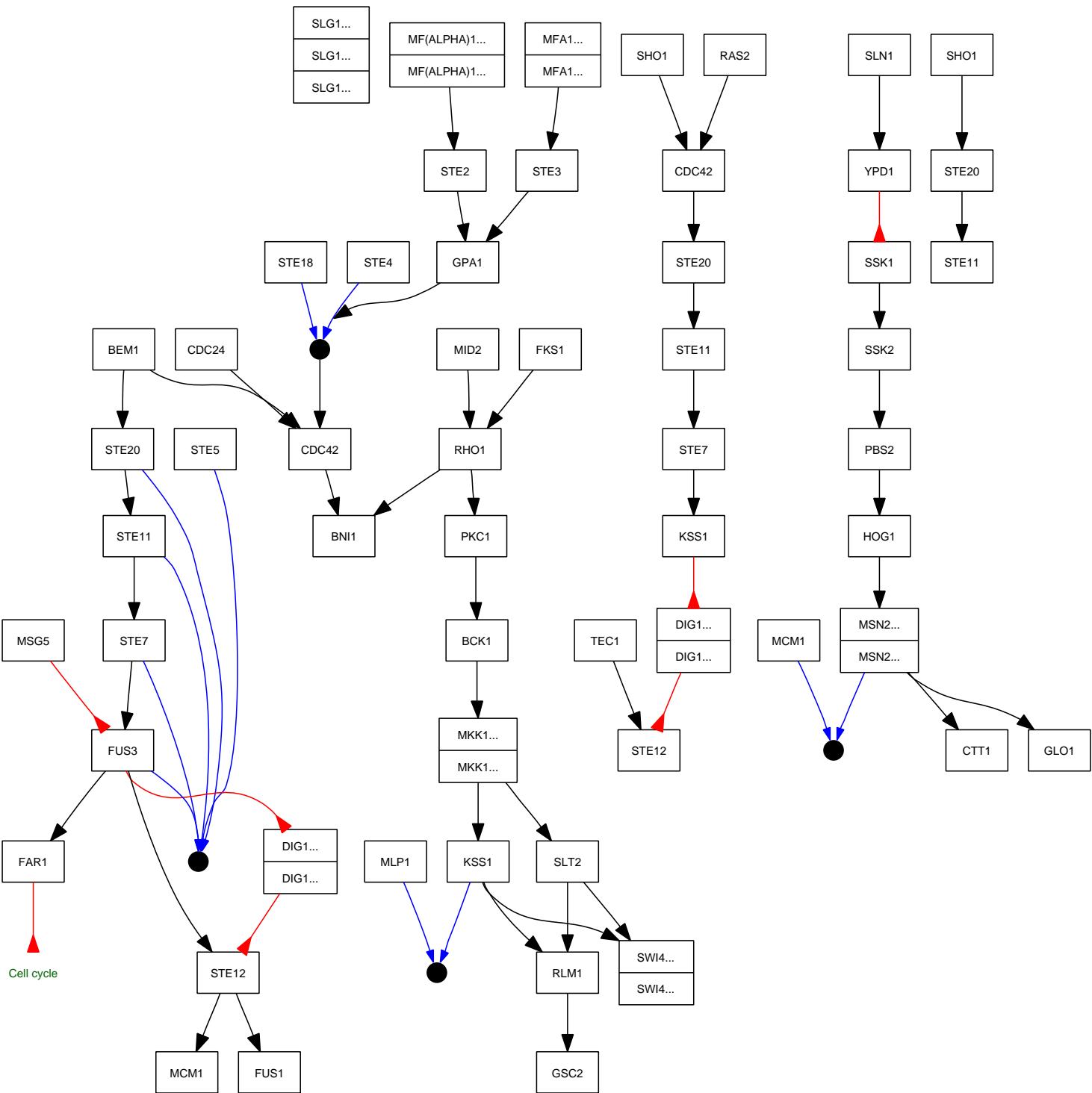
Theoretical issue

- ◆ Same as for gene set recognition
- ◆ Note: this is a much harder problem. We are not formally promising success in this aim.

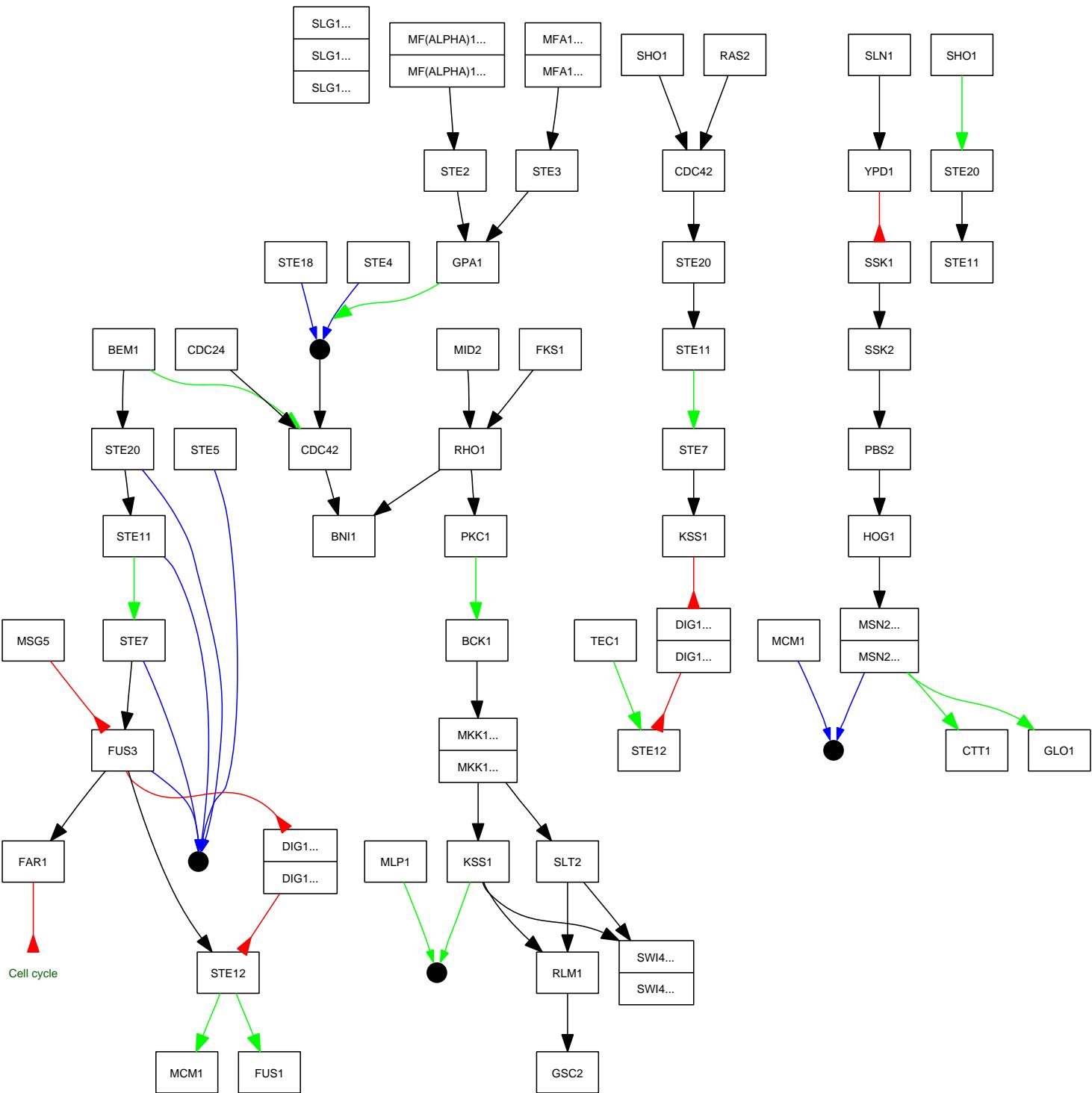
Output

- ◆ Proposed likely interconnections between genes

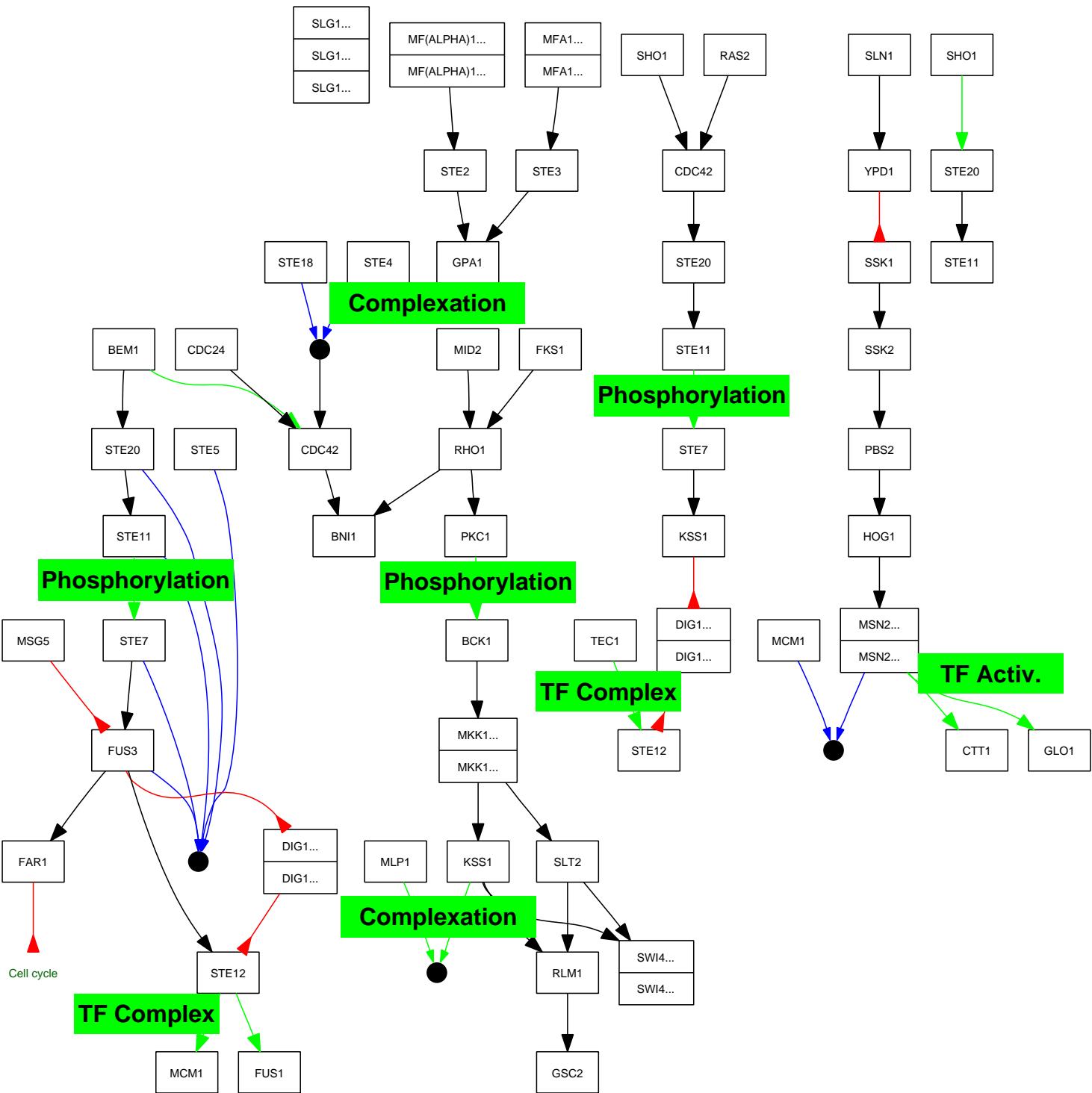
MAPK signaling pathway



MAPK signaling pathway



MAPK signaling pathway





QPACA: Conclusions

QPACA

- ◆ Interaction of experimental data with pathways
- ◆ Tight integration with Magellan
- ◆ Interoperability with pathway data exchange standards

QPACA functions

- ◆ Pathway visualization
- ◆ Pathway definition
- ◆ Statistical visualization
- ◆ Gene set recognition
- ◆ Pathway augmentation
- ◆ Pathway connectivity

Where we are

- ◆ Visualization (graphs and coloring) is solid
- ◆ Representation and language are flexible and relatively intuitive (easy to modify for interoperability)
- ◆ Gene set recognition: multiple pathways in multiple organisms show solid results for known pathways vs. random genes
- ◆ Pathway connectivity: suggestive results
- ◆ To be done: lots of stuff...



Acknowledgements

Experimental collaborators

- ◆ Albertson Lab
- ◆ Collins Lab
- ◆ Gray Lab
- ◆ Pinkel Lab
- ◆ Waldman Lab
- ◆ John Weinstein (NCI)

Jain Lab

- ◆ Lawrence Hon
- ◆ Chris Kingsley
- ◆ Tuan Pham
- ◆ Barbara Novak
- ◆ Taku Tokuyasu, PhD
- ◆ Jane Fridlyand, PhD
[Now faculty at UCSF]
- ◆ Adam Olshen, PhD
[Now faculty at Sloan-Kettering]

UCSF Biological and
Medical Informatics
(BMI) PhD Students